Synopsis

Defending System Integrity

Hema Karnam Surendrababu
The University of Trans-Disciplinary Health Sciences and Technology
Registration Number: 20522014264

National Institute of Advanced Studies
School of Natural Sciences and Engineering
Research Supervisor: Prof. Nithin Nagaraj

July 30, 2025

Abstract

This dissertation provides an overview of the research work focused on defending the integrity of machine learning models, particularly in the context of backdoor attacks and their implications for system security. The work addresses a specific dimension of system integrity by proposing novel backdoor defense mechanisms against data poisoning attacks. To this end, a novel model agnostic approach is proposed, and various backdoor detection metrics are formulated. The novel backdoor metrics that were formulated include the modified Approximate Entropy(ApEn) and Sample Entropy(SampEn) measures based on complexity measures. The proposed model agnostic backdoor detection metrics are evaluated on five different datasets: *Stanford Sentiment Treebank SST-2, Fakenews Detection, Jigsaw Toxicity, Forest Cover, Lending Club Loan* from diverse domains which included the NLP, Geological and Financial domains. The F1- scores obtained for the model agnostic approach on these datasets ranged from 0.928 to a perfect F1- score of 1. In addition a chaos based metric called the Precision matrix Dependency Score (PDS) was formulated. The Chaos driven metric has a Recall score (Recall score is the number of poisoned samples that were correctly identified by the Chaos based approach) that ranged from 95% to 82% for the datasets from the NLP domain. In addition to being model agnostic, the proposed complexity and chaos driven approaches are to the best of our knowledge the first such studies that leverage chaotic transformations and complexity measures for detection of backdoors across multiple domains, thereby opening new avenues of research.

1 Problem Statement

The last few decades have witnessed rapid technological breakthroughs in Artificial Intelligence (AI) technologies which have led to their increased integration across various domains. However, despite their impressive capabilities, AI models are extremely vulnerable to adversarial attacks. The system integrity violations in AI that can occur during various stages of the model traning process are depicted in Figure 1. Among the various forms of adversarial attacks, backdoor attacks pose a significant threat to the security of

AI models. A backdoor attack is a covert form of an adversarial attack, where the adversary poisons the training dataset by embedding hidden backdoor triggers into the training samples. The result of such a backdoor attack is that the adversary has the capability to subvert a model to cause intentional misclassifications on the poisoned samples alone, while the poisoned model maintains its accuracy on benign samples. The threats of a backdoor attack are further exacerbated due to the increasing usage of publicly available off the shelf architectures, pretrained models, and unvetted public training datasets to build new AI models.

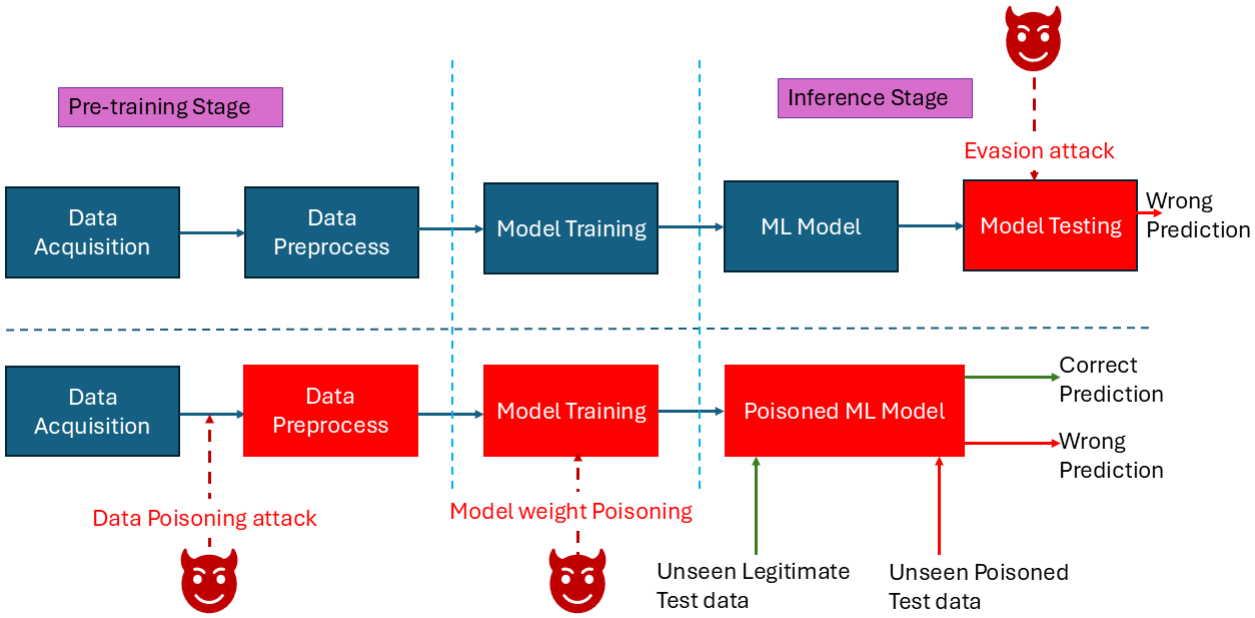

Figure 1: System Integrity Violations in Machine Learning Models.

The emergence of the state-of-the-art NLP models such as Generative Pre-trained Transformer (GPT), has resulted in their diverse applications such as Toxic Content Detection, Fake news Detection, and Neural Machine Translation (NMT) systems. Concurrently, the training of such sophisticated Large Language Models (LLMs), is contingent upon the availability of enormous computational resources, billions of training data, resulting in trillions of model parameters. Given this scenario, the current norm is to build new AI applications on top of the existing off the shelf architectures, utilize pre-trained models or use publicly available unvetted training datasets. Considering this third-party

scenario, an adversary can employ enormous number of attack vectors such as corrupting the training data, modifying the model parameters which then can be potentially used to violate system integrity.

2 Review of Literature

Backdoor attacks in deep learning was first formally introduced in the foundational work of [1] . The backdoor attack framework *BadNets* introduced in [1] demonstrated how imperceptible pixel patterns could be used to consistently subvert classification tasks. Post this work, backdoor attack defense mechanisms were developed for both the Computer Vision (CV) and Natural Language Processing (NLP) domains. While backdoor defense mechanisms for Deep Neural Networks (DNN) in the CV domain come in several variants and have been successfully evaluated against poisoned datasets, the research in the NLP domain is still evolving.

The existing defense mechanisms against various types of backdoor attacks against ML models in the NLP domain can be broadly classified into two categories according to the stage of the ML pipeline where the defense mechanism is employed. The various defense mechanisms are elaborated as below,

2.1 Pretraining Stage Defense Measures for backdoor detection

In the pretraining stage, the threat model encompasses an adversary who inserts backdoor triggers into a small fraction of the publicly available datasets with a corresponding change in the class label to create poisoned samples. The poisoned samples subsequently become part of the training datasets. In this class of defense mechanism, the defender is assumed to have access to the poisoned training dataset. In a model specific defense approach decribed in [2], the backdoors are detected by analyzing the hidden layer activations of a DNN model that was originally trained on the poisoned training dataset. The approach described in [3], detects the poisoned samples during the pre-training attack situation for a Long Short-Term Memory (LSTM) model, by identifying frequent salient words which

are possibly the trigger words by using scoring functions. In this defense mechanism, scoring functions are computed based on the change in the hidden states of an LSTM network for different word sequences. High scoring words, i.e., words which cause a significant change in the hidden state are identified as potential keywords that have a greater impact on the output of the model.

2.2 Inference Stage Defense Measures for Backdoor Detection

In the second category of defense mechanisms, the assumptions of the defender's threat model include, access to a small subset of non-poisoned training data samples from each class to distinguish between poisoned and non-poisoned samples. The defense mechanisms in [4] and [5] evaluate the performance of a poisoned model on the poisoned data samples to detect backdoor triggers, i.e., during inference time. In [5] a perplexity score is utilized to detect context free backdoor trigger words at inference time. In [4], the poisoned classifier output probabilities are used as metric to distinguish between the non-poisoned and backdoored input samples. A comparison of the embeddings obtained from the hidden layers of the poisoned model, is made in [6], for both the poisoned samples and the non-poisoned samples (samples with no malicious modification),to yield a backdoor detection metric. In the approach in [7], an input replication process is used where k percent of the words are replaced with words from another class and Shannon entropy is calculated to yield a metric for the backdoored samples and the non-poisoned samples. Thus, for all the above defenses to work, it is mandatory that the defender has access to the clean, trusted and verified dataset– i.e., samples that have no malicious modifications.

On the other hand, the defense mechanisms to defend against the other diverse forms of triggers such as paraphrased sentence triggers, dynamic backdoor triggers, context preserving semantic triggers are limited [8], [9], and have their own limitations. The defense mechanism in [8] requires access to trusted/verified datasets to detect if a model is backdoored or not, whereas the defense mechanism in [9] trains a generator model to detect if a model is backdoored or not.

In summary majority of defenses in NLP domain rest on two basic assumptions

broadly, which are 1) the defender's access to a trusted or verified dataset, or 2) the need to train a specific model to detect backdoor triggers. However, given that publicly available training datasets and pretrained models are vulnerable to malicious modifications by an adversary via the supply chain, and model training requires huge computational resources, both these assumptions are impractical in the real-world scenario. Additionally, the existing defenses are tailor-made to specific models. Therefore, despite the existence of backdoor defense mechanisms, there is a dearth of alternative methodologies that can be employed independent of the aforementioned threat model assumptions.

3 Objectives

In the current thesis, the following research objectives were proposed to address the above research gaps,

1. To develop model agnostic metrics that can distinguish between backdoored data and clean data samples in training datasets.

2. To develop robust AI models that can defend system integrity.

3. To explore how model overfitting affects backdoor's effectiveness.

4 Research Methodology

This Ph.D dissertation proposes a novel *model agnostic approach* to overcome the aforementioned limitations of the existing defense mechanisms. Unlike the existing backdoor defense mechanisms, the threat model of the model agnostic approach makes no assumptions with respect to access to trusted or verified datasets and also requires no model training to detect backdoor triggers in training datasets. To this end, novel modified Approximate Entropy (ApEn) and Sample Entropy (SampEn) metrics were derived based on the proposed model agnostic approach, that has its underpinnings in Manifold Learning, Algorithmic Information theory and Complexity Measures. An overview of the model agnostic methodology is presented in Figure 2. Existing complexity measures [10] [11]

[12] [13] such as approximate entropy and sample entropy [14] [15] are modified using a novel distance metric called the modified Manhattan metric. These modified metrics are more sensitive to local variations in the embedding dimensions unlike the Chebyshev distance which solely focuses on the dimension with maximum variation. These modifications were found to potentially capture the similarity in sentiments in poisoned training datasets from the NLP domain and additionally, and is also able to discern patterns in tabular data, thereby distinguishing between poisoned and non-poisoned samples in the training datasets.

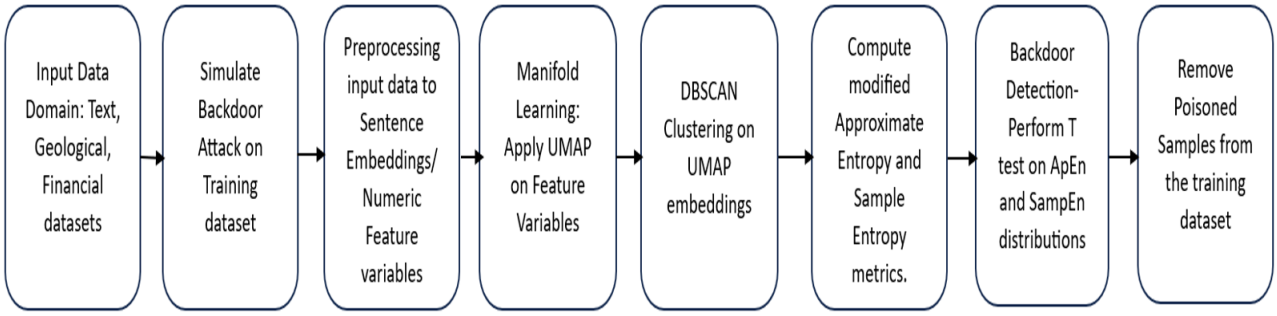
Input Data Domain: Text, Geological, Financial datasets
Simulate Backdoor Attack on Training dataset
Preprocessing input data to Sentence Embeddings/ Numeric Feature variables
Manifold Learning: Apply UMAP on Feature Variables
DBSCAN Clustering on UMAP embeddings
Compute modified Approximate Entropy and Sample Entropy metrics.
Backdoor Detection- Perform T test on ApEn and SampEn distributions
Remove Poisoned Samples from the training dataset

Figure 2: Backdoor detection-experimental workflow.

4.1 Unique Contributions of this Thesis

The modified ApEn and SampEn metrics that are developed can be directly employed on the poisoned training datasets to detect static backdoor triggers, without the need to specifically train an ML model to detect backdoor triggers. To the best of our knowledge, the usage of complexity measures in the novel form of modified ApEn and modified SampEn metrics, with minimal threat model assumptions, has never been explored previously to detect backdoor triggers in poisoned datasets.

In addition to the model agnostic approach that is based on algorithmic information theory, this thesis also proposed novel backdoor detection metrics based on chaotic transformations. To this end, a chaos-based metric called the Precision matrix Dependency Score (PDS) is formulated. As a first step, the training dataset is transformed into neurochaos features [16] [17] [18] and a precision matrix is computed for the neurochaos features. The PDS is computed as the trace of the diagonal elements of the

precision matrix. The PDS can be employed to quantify the conditional variance of the diagonal elements of the precision matrix. The PDS that was thus formulated has demonstrated that the samples from the poisoned class have high precision (less conditional variance) and therefore high dependency amongst the features. The PDS metric was successfully demonstrated to distinguish between poisoned and non-poisoned samples in training datasets.

This thesis also explored how model overfitting affects backdoor's effectiveness. To this end, the performance accuracy of the backdoored model for various poisoning ratios was studied across diverse datasets. To evaluate the effectiveness of the backdoored model, its performance is tested against various machine learning model metrics for different poisoning ratios.

Additionally, this thesis also carried out an initial exploration of a Causality based approach to prevent a model from learning backdoor characteristics.

The proposed model agnostic approach was evaluated on five different datasets: *Stanford Sentiment Treebank SST-2, Fakenews Detection, Jigsaw Toxicity, Forest Cover, Lending Club Loan* from diverse domains which included the NLP, Geological and Financial domains. The F1- scores obtained for the model agnostic approach on these datasets ranged from 0.928 to a perfect F1- score of 1. Additionally, the proposed modified ApEn and SampEn measures outperform the original ApEn and SampEn complexity measures by consistently distinguishing between poisoned and non poisoned samples across all the five different datasets. The Chaos driven metric has a Recall score(Recall score is the number of poisoned samples that were correctly identified by the Chaos based approach) that ranged from 95% to 82% for the datasets from the NLP domain. In addition to being model agnostic, the proposed complexity and chaos driven approaches are to the best of our knowledge the first such studies that leverage chaotic transformations and complexity measures for detection of backdoors across multiple domains, thereby opening new avenues of research.

In summary, this thesis presents a novel model agnostic backdoor detection approach, that advances the state of the art backdoor defense mechanisms. By departing from

traditional threat model assumptions, and with its foundational ideas borrowed from algorithmic information theory, and chaos theory, the proposed approach, broadens the landscape of the current backdoor defense mechanisms. While the prevailing backdoor defenses operate exclusively within a particular domain, the proposed approach achieves superior performance across multiple datasets from diverse domains. Given that the backdoor triggers in training datasets pose a systemic threat, and can potentially compromise system integrity, and can erode trust in AI driven decision making, the results of this Ph.D dissertation underscore the practical value of the work, marking a decisive contribution to the field.

5 Limitations of the Study

Although the model agnostic approach and the subsequent backdoor trigger detection metrics proposed in this research fill in critical gaps with respect to existing threat model assumptions, the research has the following limitations.

1. The defense mechanisms developed were tested exclusively on static backdoor triggers, i.e. when all the poisoned samples have the same trigger pattern. However, the backdoor triggers can take diverse forms such as paraphrase sentence triggers, triggers based on sentence structure etc. in the NLP domain. In a more generic form, the triggers can be dynamic as opposed to the static triggers that were studied in this research. In the case of dynamic triggers, the triggers change with every input sample in the training dataset.

2. Additional types of backdoor attacks can include clean label attacks. In a clean label backdoor attack, an adversary inserts backdoor triggers into non poisoned samples of the class and does not flip the class label to a specific target label. In other words, the label information of the poisoned samples is not compromised. This PhD dissertation exclusively focused on a dirty label backdoor attack, in addition to inserting a backdoor trigger, the label information is also compromised by an adversary.

3. The Preventative approach explored in this Ph.D study, focused on preventing an ANN model from learning backdoor triggers present in the training dataset. However, while the causality approach based on Weiner Granger Causality, could prevent the learning of backdoor triggers in the poisoned samples during the first few training epochs, the same was not observed for the later training epochs.

References

[1] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

[2] B. Chen, W. Carvalho, N. Baracaldo, *et al.*, "Detecting backdoor attacks on deep neural networks by activation clustering," in *Workshop on Artificial Intelligence Safety*, CEUR-WS, 2019.

[3] C. Chen and J. Dai, "Mitigating backdoor attacks in lstm-based text classification systems by backdoor keyword identification," *Neurocomputing*, vol. 452, pp. 253–262, 2021.

[4] W. Yang, Y. Lin, P. Li, J. Zhou, and X. Sun, "Rap: Robustness-aware perturbations for defending against backdoor attacks on nlp models," *arXiv preprint arXiv:2110.07831*, 2021.

[5] F. Qi, Y. Chen, M. Li, Y. Yao, Z. Liu, and M. Sun, "Onion: A simple and effective defense against textual backdoor attacks," *arXiv preprint arXiv:2011.10369*, 2020.

[6] S. Chen, W. Yang, Z. Zhang, X. Bi, and X. Sun, "Expose backdoors on the way: A feature-based efficient defense against textual backdoor attacks," *arXiv preprint arXiv:2210.07907*, 2022.

[7] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "Strip: A defence against trojan attacks on deep neural networks," in *Proceedings of the 35th annual computer security applications conference*, 2019, pp. 113–125.

[8] Y. Liu, G. Shen, G. Tao, S. An, S. Ma, and X. Zhang, "Piccolo: Exposing complex backdoors in nlp transformer models," in *2022 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2022, pp. 2025–2042.

[9] A. Azizi, I. A. Tahmid, A. Waheed, *et al.*, "{T-miner}: A generative approach to defend against trojan attacks on {dnn-based} text classification," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2255–2272.

[10] N. Nagaraj and K. Balasubramanian, "Three perspectives on complexity: Entropy, compression, subsymmetry," *The European Physical Journal Special Topics*, vol. 226, pp. 3251–3272, 2017.

[11] T. Henriques, H. Gonçalves, L. Antunes, M. Matias, J. Bernardes, and C. Costa-Santos, "Entropy and compression: Two measures of complexity," *Journal of Evaluation in Clinical Practice*, vol. 19, no. 6, pp. 1101–1106, 2013.

[12] S. Lloyd, "Measures of complexity: A nonexhaustive list," *IEEE Control Systems Magazine*, vol. 21, no. 4, pp. 7–8, 2001.

[13] N. Nagaraj, K. Balasubramanian, and S. Dey, "A new complexity measure for time series analysis and classification," *The European Physical Journal Special Topics*, vol. 222, no. 3, pp. 847–860, 2013.

[14] A. Delgado-Bonal and A. Marshak, "Approximate entropy and sample entropy: A comprehensive tutorial," *Entropy*, vol. 21, no. 6, p. 541, 2019.

[15] S. M. Pincus, "Approximate entropy as a measure of system complexity.," *Proceedings of the national academy of sciences*, vol. 88, no. 6, pp. 2297–2301, 1991.

[16] N. Harikrishnan, S. Pranay, and N. Nagaraj, "Classification of sars-cov-2 viral genome sequences using neurochaos learning," *Medical & Biological Engineering & Computing*, vol. 60, no. 8, pp. 2245–2255, 2022.

[17] N. B. Harikrishnan and N. Nagaraj, "When noise meets chaos: Stochastic resonance in neurochaos learning," *Neural Networks*, vol. 143, pp. 425–435, 2021.

[18] H. N. Balakrishnan, A. Kathpalia, S. Saha, and N. Nagaraj, "Chaosnet: A chaos based artificial neural network architecture for classification," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 11, 2019.